A Critical Assessment of a Method for Detecting Diachronic Meaning Shifts: Lessons Learnt from Experiments on Dutch

Hessel Haagsma* HESSEL.HAAGSMA@RUG.NL
Malvina Nissim* M.NISSIM@RUG.NL

* *CLCG, University of Groningen, Oude Kijk in 't Jatstraat 26, Groningen, The Netherlands*

Abstract

Automatically detecting shifts of meaning over time is desirable for Natural Language Processing tasks as well as research in the Digital Humanities. We train diachronic word embeddings on Dutch newspaper data and compare representations of the same terms from different times to each other. The aim is verifying whether such comparison can highlight the emergence of a new (figurative) meaning for a given term. The most interesting outcome of this experiment is methodological: while in some cases we observe that this method is efficient, as it has been shown to be in previous work on Italian, we also observe that in many other cases results are not what we would expect. This leads us to an in-depth analysis of interfering aspects and to a discussion of methodological choices, towards the development of what we believe is a timely and needed roadmap for developing diachronic studies on meaning shifts that rely on distributed representations.

1. Introduction

The meanings of words are known to change over time, and the rise of new figurative uses is a fundamental driver of this change (Haser 2003). Detecting meaning shifts is beneficial to any NLP task which relies on lexical meaning resources, as well as for research in lexical semantics. More concretely, the aim of this line of research is to be able to automatically detect shifts in word meaning in a completely unsupervised fashion from raw text, often relying on distributional semantic representations.

The idea behind exploiting such representations is that the semantic space in which a given word is located might change over time, if the meaning of the word itself changes at least partially, such as with the emergence of a new meaning. For example, one could expect that the location in semantic space of the word 'mouse' changed drastically after the computer tool's invention by Douglas Engelbart in the 1960's, thereafter called 'mouse'. Conversely, the semantic space of a concept has been shown to stay pretty much the same across time, becoming a reliable indicator of that very same concept, even when the word used to express it changed in time. An example of this are the words 'Walkman' and 'iPod' expressing the 'portable music player'-concept (Jatowt and Duh 2014, Kenter et al. 2015). Indeed, experiments in detecting diachronic semantic shifts have been done in both directions: detecting whether a word might gain new (figurative) meanings over time, and detecting whether a meaning might become expressed by new words over time.

Building on the assumptions that (i) the semantic space of a word can be captured via distributed semantic representations in the form of word embeddings, (ii) the embeddings of the same word can become substantially dissimilar across time if a new meaning emerges, and (iii) this can be particularly true in the case of metaphoric shifts as the conceptual fields of occurrence differ a lot, Del Tredici et al. (2016) propose a method for comparing and interpreting the spaces of five Italian words that are known, based on dictionary information, to have undergone a metaphoric shift. In their experiments, they observe that a drop in similarity of the embeddings of the same word over time often coincides with the introduction of a new figurative meaning in a reference Italian dictionary.

In order to further test the hypothesis of a correlation between a change in embeddings and the introduction of a new meaning, we replicate Del Tredici et al.'s experiment on Dutch data. By doing so, we obtain a picture of meaning shift for a selection of Dutch words that is substantially fuzzier than was observed for Italian data. More specifically, we stumble across several methodological aspects that we believe deserve proper attention, before we can glean sensible interpretations of the data. We believe that what we observe can lead to improvements in the field of the automatic detection of diachronic meaning shifts, and we establish a set of directions towards this end in our discussion.

Contributions Therefore, the aim of this paper is twofold: (i) presenting results of a study where the semantic space of some pre-selected Dutch words is compared across time, so as to spot a potential dissimilarity of a word to itself that can be explained in terms of meaning shift (especially figurative); and, most importantly, (ii) critically assessing all of the factors that contribute to what we observe and how results can be interpreted, with particular attention to frequency, polysemy, and other confounding factors. Following such observations, we offer an overview of what we think are lessons learnt from this work, which can be turned into recommendations for more robust, reliable, and scalable practices in setting up and evaluating meaning shift detection experiments based on distributional semantic representations.

2. Approach

2.1 Data

We collect a corpus of Dutch newspaper articles by scraping the web pages of *Trouw* and *Volkskrant*. The corpus covers the time period 1994–2016, which is all that is available in the online archives of both newspapers. All articles were tokenized using Elephant (Evang et al. 2013), resulting in a corpus of 800 million tokens, of which 485 million are from *Volkskrant* texts, and 315 million tokens are from *Trouw* texts. No lemmatization is performed. The corpus is organized in year-slices per newspaper, e.g. all text from *Trouw* in 1996 forms one subcorpus, text from *Volkskrant* in 2005 another, and so on. The sizes of the different subcorpora diverge significantly, ranging from 4.3 million (Volkskrant-1994) to 35.6 million (Volkskrant-2013).

2.2 Training Embeddings

We aim to track semantic change by measuring self-similarity for a given word, across time. To do this, we need a set of diachronic embeddings, i.e. one set of embeddings per year-slice. We train these diachronic embeddings by broadly following the method presented in Del Tredici et al. (2016).[1]

In this method, diachronic embeddings are not made by training separate embeddings sets on different year-slices and aligning them. Rather, first embeddings are trained on the most recent year (2016, in our case); then the obtained embeddings are used as the initialization for the embeddings of the year prior, and then trained on the data of that year slice. This is repeated until the data of the least recent year (1994, in our case) has been incorporated. This removes the need for alignment, since all embeddings, regardless of year, exist in the same vector space. This, in turn, allows for direct comparison of embeddings year-to-year, which is our goal. The reason for training in a 'backwards' fashion, rather than from the oldest year-slice to the newest one is the same as in Del Tredici et al. (2016): we expect the words with novel meanings to occur with sufficient frequency in the most recent year-slice, but not necessarily in the oldest year-slice.

As for the practical implementation, we use the `gensim` Python library (Řehůřek and Sojka 2010) to train our embeddings using the `word2vec` skip-gram method (Mikolov et al. 2013). We use the two newspaper corpora separately, that is, we train separate sets of diachronic embeddings for each newspaper. We initialize the vocabulary on the set of all words from all years of a given newspaper

1. Source code for this work available at `https://github.com/hslh/diachronic-meaning-clin`.

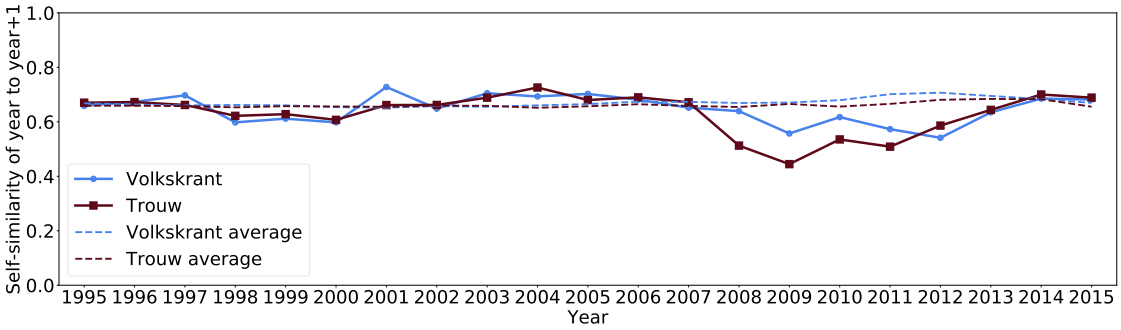

Figure 1: Cosine self-similarity by year for *terreur*, 'terror'. Note the clear self-similarity dip in 2007–2014.

corpus. Words with a total frequency below 30 are ignored. For the skip-gram modelling, hierarchical softmax is used, a learning rate of 0.01, an embedding dimensionality of 200, a window size of 5, and 20 iterations of training over each year-slice.

2.3 Selection of Words

We use a pre-selected list of words to investigate for new metaphorical meanings. This list was composed from two sources: suggestions of possible candidates by the authors, and from the Van Dale dictionary, the editors of which, upon request, kindly provided us with a full list of words which gained a new, figurative meaning in the dictionary at some point between 2000 and 2016. The full list contains 321 words, most of which are highly infrequent and have very marginal new meanings. Over 95% of these words were provided by Van Dale.

Since our corpus-based method only works for words which occur frequently enough, we select a subset of these 321 words for investigation, using the following criterion: the word has to have a minimum average frequency of 10 words per million across all year-slices in both newspapers. This results in a final selection of 23 words.[2] The list comprises nouns, verbs, and adjectives, as among our choice criteria of word selection we did not set any constraint for part-of-speech. This has the advantage of exploring meaning shift beyond just nouns, as it's usually done. Note though that because our corpora are not pos-tagged, we cannot control for part-of-speech in the matched data, so that the word *laag*, 'layer, low, stratum', for example, which was included in the Van Dale's list in its noun form, will be found and considered in all of its occurrences, adjectives included.

3. Analysis of Results

In this section we review the results we obtain by first showing what can be observed in the data, and then by discussing several cases, including both positive and negative examples. In addition, we offer explanations for the dishomogeneous nature of the results, illustrating them with examples.

In Figure 1, and other figures, four values are plotted across the time span under considerations. The solid lines represent the similarity of a given word to itself (here, *terreur*, 'terror') in the Volkskrant corpus (light blue line) and the Trouw corpus (dark red line). The dashed lines show the

2. These words are, in order of frequency: *laag*, *vertrek*, *bieden*, *podium*, *ziek*, *breken*, *bodem*, *bewegen*, *nazi*, *hoofdrol*, *keihard*, *architectuur*, *terreur*, *recept*, *management*, *slim*, *monument*, *dodelijk*, *vet*, *platform*, *dictatuur*, *bloedbad*, *formaat*. Translations are provided in the text when a specific word is used.

average self-similarity across time for the 5000 most frequent words in Dutch as obtained from the Lassy Small corpus (van Noord et al. 2013).

According to our hypothesis, and as partly shown in (Del Tredici et al. 2016), the introduction of a new (metaphorical) usage of a word should cause a change of contexts for that word, which should show up as a drop in self-similarity in the plot at a certain point in time. In Figure 1, such a drop can clearly be observed around the years 2007–2014 for both newspapers.

The drop is observed simply by manual inspection of the plots. What is shown in Figure 1 is an optimal case with a very clear pattern, which is not what we observe for most other sampled words. In Figures 2 to 6 we present a representative sample of unclear and non-ideal cases, which will serve to highlight a set of confounding factors and representation issues that we discuss in depth in what follows, as we believe they influence the observations we derive and their analysis. These factors have both to do with intrinsic properties of language and this phenomenon in particular (such as frequency variation and polysemy), as well as with artefacts that are introduced by our modelling choices (such as the corpus we use and the representation method we adopt).

3.1 Influence of frequency

Frequency affects the settings and results of our experiments in at least two ways.

First, absolute frequency in the whole dataset affects the stability of the representation, mainly in that flow frequency will make the representations extremely unreliable and therefore little informative. An example of this is the word *leermeester*[3], 'master, guru, teacher', which is plotted in Figure 2. As we can see, self-similarity is always below average, and shows large, random fluctuations, which we take to be indicators of an unstable representation – due to there not being enough instances of *leermeester* in the corpus for its representation to stabilise.

Second, heavy fluctuations in frequency across the time span cause instability in the representation of a word, so that drops in self-similarity can potentially be explained as a by-product of frequency peaks, rather than as an indication of meaning shift. This is illustrated well by the example of *architectuur*[4], 'architecture', in Figure 3. In Figure 4 we show the frequency distribution across the time slices for the same word. It is rather straightforward to observe that the drop in self-similarity in the Trouw around 2010-2012 has a direct correspondence in a frequency peak in 2011. One might argue that the emergence of a new meaning can also prompt an increase in frequency, as there are more contexts in which to use the word, but it is unclear how such behaviour should be accounted for in modelling diachronic shift using currently existing methods. (see also Section 3.3 on sense change and topic change).

3.2 Influence of polysemy

The use of distributional semantic models for the detection of meaning shifts relies on the assumption that the emergence of a new sense will mean that it will be used in a different context, which deviates from the 'default' contexts the word was previously used in. However, this assumption becomes a lot more unreliable when we are dealing with highly polysemous words. In that case, self-similarity across years will be based on a representation that covers a multitude of senses, so that the emergence of a new one will only shift the representation very slightly, perhaps too little to be detected.

In their work, Del Tredici et al. (2016) clearly specify that in the choice of words to analyse they took into account polysemy, and their terms are either monosemous or minimally polysemous. We do not control for this in our word selection. On the contrary, we select our words based on a frequency threshold, which exacerbates the polysemy issue, as more frequent words tend to also be more polysemous (Zipf 1949). Double influences of high polysemy and high frequency makes

3. The new sense consists of a metaphoric shift from *leermeester* referring to people only, to being able to refer to events and experiences as well.
4. The new sense consists of an extension of meaning from only concerning physical structures to concerning all kinds of physical and abstract frameworks.

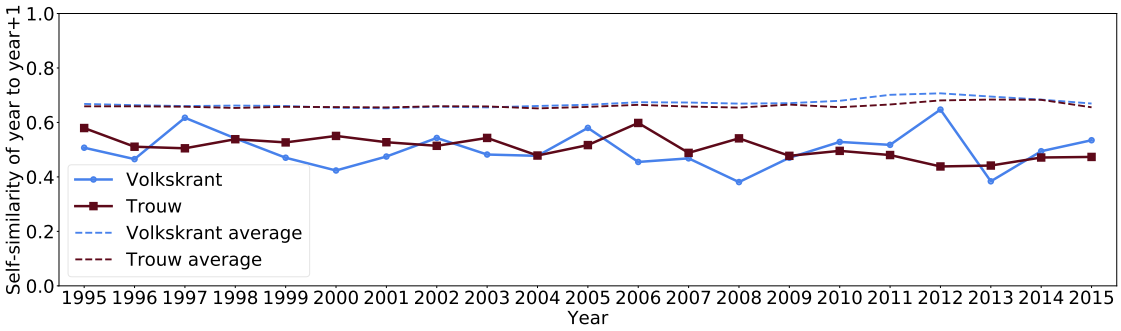

Figure 2: Cosine self-similarity by year for *leermeester*, 'master, guru, teacher'. Illustrates the effect of low overall frequency on self-similarity.

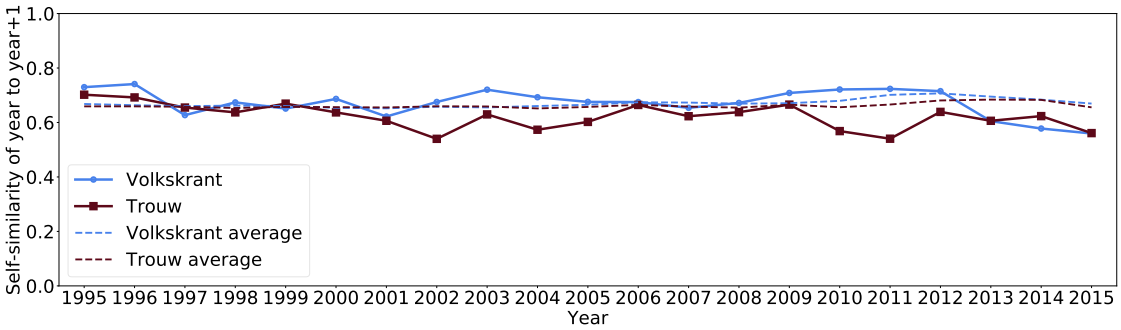

Figure 3: Cosine self-similarity by year for *architectuur*, 'architecture'. Illustrates the effect of frequency fluctuations on self-similarity.

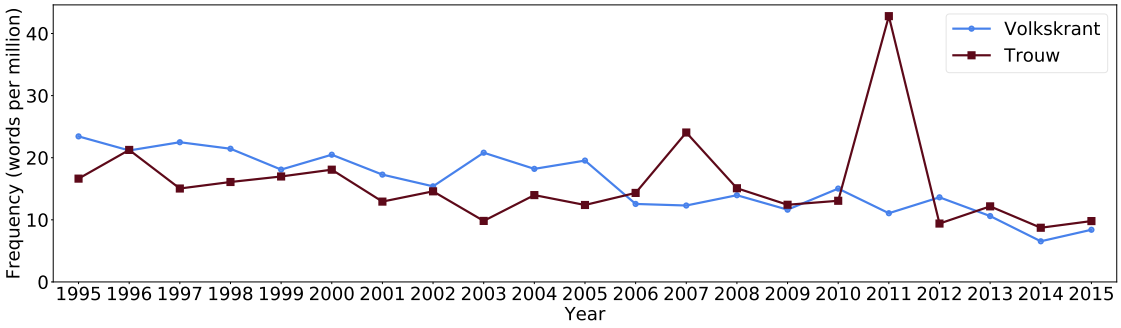

Figure 4: Frequency per million words of *architectuur*, 'architecture'. Note the frequency peak in Trouw in 2011.

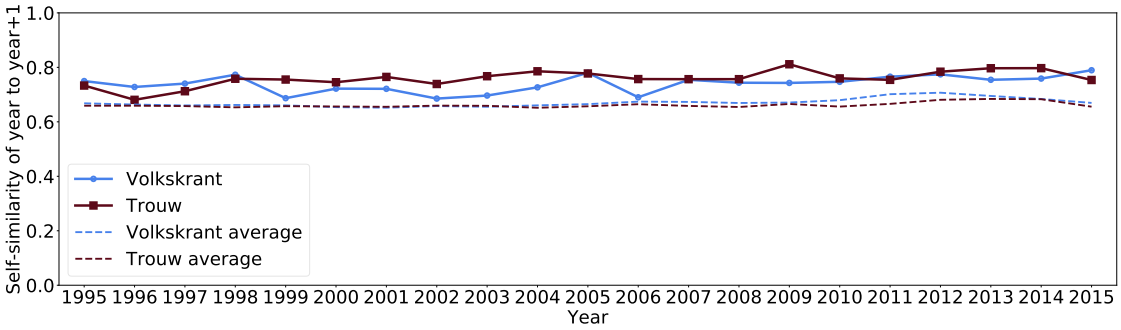

Figure 5: Cosine self-similarity by year for *laag*, 'layer, low, stratum'. Highly polysemous word, has 7 senses in Cornetto (Vossen et al. 2013), 23 in Van Dale.

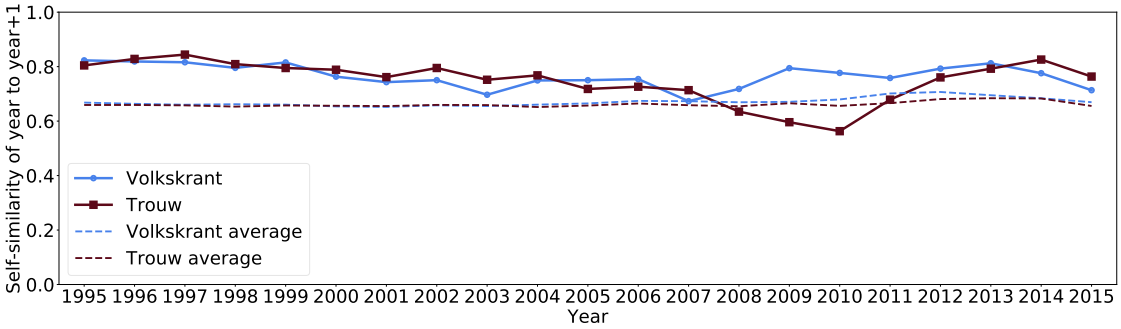

Figure 6: Cosine self-similarity by year for *nazi*, 'nazi'. Illustrates the discrepancy between the *Volkskrant* and *Trouw* corpora.

the analysis conducted under our method quite unreliable, as these two factors tend to blur any significant meaning changes observable through self-similarity. This shows clearly in Figure 5, where we show the behaviour of the word *laag*[5], 'layer, low, stratum'. Here, we see a high, above-average self-similarity, which is very stable across time and corpora. Clearly, if the new sense of *laag* is used in the corpus, it does not influence the representation enough to clearly show up in the plot.

3.3 Sense Change and Topic Change

Without manual inspection, it is hard to know whether a change in self-similarity (i.e. a change in the word's contexts) is caused by a new sense or a shift in sense, or by the use of the word in the same sense, but mainly on a different topic. A related problem is that it is sometimes hard to say whether a shift to a main use regarding a new topic constitutes a new sense or not.

This exacerbated by the fact that our corpus consists of newswire text, which is very much driven by real-world events and popular topics. As such, it could be that the word *bodem* shows a significant drop in self-similarity, not because it acquires a new sense, but because it is was mainly used when

5. The novel sense is roughly equivalent to *betekenislaag*, 'layer of meaning', as used when talking about a work of art.

talking about finance at one point (*de bodem van de schatkist* - 'the bottom of the treasury'), and about earthquake trouble later (*de Groninger bodem* - 'the soil of Groningen').

3.4 Differences between Corpora

We make use of two very similar, but distinct corpora: *Trouw* and *Volkskrant*. Because they are so similar — they are of the same genre, time period, and country, are both considered 'quality newspapers', and likely deal with the same topics — we expect to see the same semantic changes in both corpora. Potentially, the current study would allow us to investigate this hypothesis, that similar corpora show similar semantic changes, or, alternatively, find points where they differ, and analyse those differences in more detail.

In practice, however, things are less ideal, and more complex. Looking at the self-similarity over time for *nazi* 'nazi' in Figure 6, for example, we see a nice, clear dip in self-similarity, around the time where we would expect it to be. Unfortunately, that is only the case for the *Trouw* corpus. In the *Volkskrant* corpus we see no clear trends or effects, at all. This leaves us with three possible explanations:

1. Ideally, the plot represents the true situation, and there is no change in *Volkskrant*. There is a true difference between the two newspapers, e.g. due to different audiences or style conventions, and manual analysis should be able to confirm this.

2. Less ideally, it turns out that the same semantic shift did occur in *Volkskrant* as well, but because of the limitations of our method, we do not see it.

3. Worst case, the dip we see in the *Trouw* plot is an artefact, and there is no shift in either of the corpora.

We cannot tell which of these possible explanations is true, unless we manually analyse the corpora, which is exactly what we try to circumvent by using an automatic method for detection of meaning shifts. If this was just the case for *nazi*, this would not cause a major problem, but we see the same for other words, such as *management*, *monument*, and *dodelijk*.[6] This poses a serious problem for the drawing of strong conclusions from the analyses for these words.

3.5 Compounding in Dutch

A distinctive feature of Dutch, when compared to English, is the prevalence of noun-noun compounds. Of course, noun-noun compounds exist in English as well, but Dutch orthography dictates that they are always written as a single word, whereas they tend to be written separately in English. This is a small difference, but it has a potentially large influence on our method. All embeddings are created on the token-level, and our tokenization procedure does not involve compound splitting. Therefore, when applied to English, the embedding for *nazi* would be influenced by the occurrence of 'grammar nazi' in the corpus, whereas for Dutch, the embedding for 'nazi' does not take into account an equivalent occurrence of *taalnazi*.

Although noun-noun compounds are frequent, we might choose to simply ignore them, given that we have a large enough amount of data (similar to how we do not take into account inflectional variation). However, we know that compounding is one of the most important mechanisms for the creation of neologisms, and thus, for novel sense creation (Booij and van Santen 1998, p. 147). This seems to be especially true for the novelty- and topic-driven writing one encounters in news media. This is reflected in our newswire corpora, where it is easy to find compounding examples for words from our word list, as illustrated by Examples (1) to (3). If our goal is to detect novel senses, we lose a significant amount of information by not dealing with compounds.

6. Plots for these words not shown here, but available at `https://github.com/hslh/diachronic-meaning-clin`.

(1) Met enige regelmaat blijkt het spel krachtvoer voor taalnazi's.
With some regularity turn.out.to.be the game concentrate for grammar.nazis
'Every now and then, the game show is a target for grammar nazis' (Volkskrant, 2014)

(2) [..] in Ede en veel andere provinciestadjes hebben ze vooral oog voor de
[..] in Ede and many other provincial.towns have they mainly eye for the
hondenpoep- en parkeerautomatenterreur .
dog.poo and parking.ticket.machine.terror.
'[..] In Ede, and many other country towns, the main concerns are the terrors of dog poo and parking ticket machines.' (Trouw, 2007)

(3) We hebben een soort mediadictatuur in Nederland.
We have a kind media.dictatorship in Netherlands.
'We have a kind of dictatorship of the media in the Netherlands.' (Trouw, 2010)

Although compound splitting is not a trivial task, we could implement it as an additional preprocessing step, and keep the rest of the method as is. Still, this would raise a rather crucial issue which has to do with the fact that instead of one vector we would have two (one for *taal* and one for *nazi*). While a promising solution has been proposed for dealing with adjective-noun composition (Baroni and Zamparelli 2010), it would still remain to be explored whether a similar approach would fit our cases at hand, especially considering the focus on novel sense emergence. In this sense, splitting compounds would give rise to an even more fundamental issue regarding what the novel sense should be attached to: in *taalnazi*, is there a novel sense of *nazi* that we should detect, or is this just a new word expressing a new meaning and should it be treated as such? Moreover, compound words are words in their own right, and can undergo meaning shifts similar to simplex words. If we split compounds in preprocessing, we lose the ability to investigate the semantic evolution of compound words.

4. Lessons learnt

Based on Section 3.1, we summarise our observations in terms of points to be taken into account when designing an experiment focused on detecting meaning shifts using distributed representations. While we do not provide definite answers or methods on how to tackle some of the intrinsic problems this kind of experiments involves, we aim at highlighting them so as to incorporate them in the design phase of a new study.

Thus, the question guiding this summary is the following: assuming that novel sense detection is useful and interesting, from a natural language processing, but also from a digital humanities and social perspective, what are the requirements on on the performance of the method for it to be useful, reliable, and scalable?

In order for automatic detection of meaning shifts across time to scale up we need to move towards a bottom-up rather than a top-down evaluation approach. In doing so, the experiments we have described point to several aspects that need to be taken into consideration. While it is far from trivial to disentangle phenomena and factors, as they are all correlated and interact with each other, we want to offer a summary here.

First, frequency is a crucial factor, both in the overall level of occurrence, as well as in terms of fluctuation across years, as it can make representations unreliable.

Second, polysemy has to be accounted for, and decisions must be made on how it is detected and quantified (external resources or directly from the data). What senses are there? Which ones are used/implied? How to account for that in diachronic experiments?

Third, because we observe variation across different datasets (which are not even that different in our case), drawing claims should be probably limited to the very specific context they come from

or should be further tested on a range of different datasets in order to be able to generalise to a wider level.

Fourth, a change in context or usage does not necessarily correspond to a change in meaning. This is especially true for highly polysemous words. This observation concerns the underlying hypothesis on which the whole work is built, but we believe it is a factor to be taken into account rather than an obstacle.

Lastly, language-specific issues must be considered. For example, in Dutch and other Germanic languages compounding is a powerful tool to encapsulate new meanings of the words in a compound, but this escapes an analysis that is purely token-based. Other interesting examples would come from morphologically rich languages, where only certain variations might undergo a meaning shift (e.g. a female word form vs. a male word form), while others do not.

5. Previous Work

The main lines along which we want to compare our contribution to existing work are related to the issues we have discussed in the previous section. Specifically, we discuss similar and different representation methods, corpus size, the problem of context vs. sense discrepancy, and the influence of frequency and polysemy. Since we attempt to reproduce only a single method, namely that of Del Tredici et al. (2016), these issues might be partially specific to the current method, and some of them have already been noted or addressed in previous work. By taking this into account, we can get a clearer view of which challenges remain for this task.

The method used to represent word meaning in this work and in Del Tredici et al. (2016), training word embeddings diachronically by initializing the embeddings for year n on those trained on year *n-1* or *n+1*, is based on the work by Kim et al. (2014). This is a good way of getting embeddings trained on different subcorpora (here, year-slices) to exist in the same embedding space. Moreover, it is intuitively elegant, since the meaning shift in real time is recreated by an embedding shift during training time.

However, other ways of integrating embeddings from different time-period defined subcorpora exist. Hamilton et al. (2016), for example, train embeddings separately for each time-slice, and use Procrustes alignment to enable cross-time similarity calculations. A second alternative is to circumvent the necessity of having embeddings in the same vector space altogether, as is done by Kenter et al. (2015). They use a set of seed words to track the vocabulary used to describe a certain concept through time, and do this across independent vector spaces by looking at the nearest neighbours of the seed words. This method could easily be applied to the detection of meaning shifts, e.g. by looking at the nearest neighbours of the word of interest.

Another notable difference between this work and previous work is the choice of corpus, corpus size, and corpus language. We use two Dutch newspaper corpora, of 315M and 485M words, spanning a timeframe of approximately 20 years. This is, of course, a limited view of the Dutch language, and might contribute to some of the issues we face. These issues, in turn, might be resolved by using a bigger corpus, spanning a larger period of time, and containing more variation.

There is previous work on Dutch using a corpus of various newspapers spanning a longer period (1950–1990), and containing approximately 2B words (Kenter et al. 2015), whereas Hamilton et al. (2016) experiment with 3 different corpora for English, containing different genres, and spanning a 200-year time period, consisting of 410M to 850B words. When comparing the different English corpora, Hamilton et al. note that the biggest corpus also shows the clearest, least variable, and most accurate results. In addition, they provide a multilingual view, experimenting on large corpora of German, French, and Chinese. Still, most work considers only English text, and often newswire, and more work on comparing meaning shift detection methods across languages and genres has to be done.

Most, if not all, work on detecting meaning shifts make use of some sort of distributional meaning representations, and this carries with it the assumption that the context in which a word is used

defines the meaning of the word. The fact that this assumption constitutes a potential source of error is rarely discussed explicitly. Gulordava and Baroni (2011) provide two clear examples of false positives thrown up by their system, cases where the context of usage changed quite drastically, but the sense of the word itself was judged to not have changed significantly. For example, they find that their system detects a shift in the discourse around *sleep* from mainly positive to mainly negative, focused on sleep problems.

Gulordava and Baroni suggest that this might be a useful preliminary indication of meaning change, but in the meantime, distinguishing sense and context change remains a problem for meaning shift detection. This is corroborated by Hamilton et al. (2016), who note that a significant number of the false positives of their system fall in the same category of discourse change without sense change. For example, they suggest that a context change for the word *male* is caused by an increased use of the word in the discourse around gender equality, rather than by the word acquiring a new sense.

Besides the issue of sense vs. context change, we are also not the first to notice the problems of frequency and polysemy. For example, Kim et al. (2014) note that their method only works for words which have a high enough frequency over the whole time period, and suggest using a measure that incorporates both similarity and frequency. More insights on how frequency and meaning shifts interact are provided by Gulordava and Baroni (2011), who find that words with a decreasing frequency over time are less likely to change, whereas words which acquire a new meaning (e.g. technology-related words like *disk*, *address*), are more likely to see an increase in frequency of use. Hamilton et al. (2016), in turn, explicitly set out to quantify the relation between frequency and rate of semantic change (and polysemy), finding that high-frequency words are less likely to change, and that polysemous words are more likely to see semantic change (controlling for frequency).

Gulordava and Baroni look at change in frequency over time, while Hamilton et al. look at average frequency over time, but neither looks at the combination of both: that is, how to deal with the interplay between overall frequency, frequency variation (including more complex variations than simple increases or decreases over time), and detection of semantic shifts.

6. Future Work

In addition to pointing out the difficulties that work on automatic detection of meaning shifts has to face, we aim to contribute to the future development of the field by highlighting a series of possible improvements and goals.

The first of these concerns evaluation. Here, as in other works (Frermann and Lapata 2016, Del Tredici et al. 2016), we use a form of top-down evaluation, where a set of words with known meaning shifts is selected, and it is assessed whether the detection method can highlight these shifts in a particular corpus. The main problem with this approach is that while they are attested in a dictionary, it is unknown a priori whether these shifts actually occur in the corpus (and if so, with what frequency), so that we cannot assess whether our method really detects all the expected shifts, or whether they were never there in the first place. Additionally, while a top-down approach might serve well those who want to investigate some specific terms, for example in the digital humanities field, it cannot scale up towards a larger and more comprehensive study of meaning shift, even for a given language within a given time frame.

A better alternative is a bottom-up evaluation approach, where shifts are detected automatically from the data via some developed measure, and only the detected shifts are then evaluated (e.g. Hamilton et al. (2016)). Thus, we go from asking '*Can we spot the expected meaning shifts in the data?*' to '*Are the meaning shifts spotted in the data actual meaning shifts?*'. A downside of this bottom-up evaluation is that it needs to be done manually, by relatively expert annotators, or resource-based, e.g. using dictionaries from different time periods. As such, in order to do this evaluation in a consistent way that is reproducible and comparable to related work, a clear definition of what is and what is not a meaning shift is required. A large corpus annotated for this phenomenon is obviously costly to create (Kenter et al. 2015), and can never be truly exhaustive.

Having to make do with manual inspection of potential shifts, a useful tool, and an overall improvement, would be a system that can provide examples of pre- and post-shift instances of a word for manual inspection, both during development and evaluation. Ideally, these would be the examples that are most typical of the unshifted and shifted meaning, to prevent the necessity of inspecting large numbers of examples. This would aid researchers in developing systems, annotators in evaluating detected meaning shifts, and overall, allow for assessment of the underlying hypothesis that sense and context changes are always concurrent.

Next to evaluation, dealing with polysemy (and homonymy) is an important issue. For words with multiple meanings, it can be hard to distinguish whether a presumed meaning shift is due to a shift in the distribution of existing senses (i.e. going from having sense-1 being dominant to sense-2 becoming much more frequent), or due to a true meaning shift or the rise of a novel sense. Moreover, by combining all senses of a word in a single representation, the granularity of the representation becomes lower. This is the case, for example, where the addition of a sense to a word with 5 or more existing senses might not cause the overall averaged representation to shift significantly. On the other hand, by splitting representations to some extent, we can detect smaller shifts, for example when a shift in the meaning of the noun *book* is not obscured by the meaning of the verb *book*.

This would of course, require word sense disambiguation, which is a largely unsolved task in NLP. Luckily, for this to yield benefits, disambiguation at the fine-grained sense-level is not required. Making coarse-grained distinctions and/or simply distinguishing between homonyms could already increase significantly the granularity of representations. Alternatively, the disambiguation step can be integrated in the representations, by making use of sense-aware embeddings as done for example by Cheng and Kartsaklis (2015) and Iacobacci et al. (2015).

On a more concrete level, the development of meaning shift detection methods would benefit from more comparability. Most existing work makes use of different corpora and evaluation methods, making it difficult to clearly assess which methods work better, and what the effect of corpus size and the use of certain similarity metrics is. More detailed studies need to be carried out, focusing on a direct comparison of different representation methods, e.g. ones using re-initialization, alignment by rotation, or nearest-neighbour-based methods for comparing different vector spaces.

As a final recommendation for future work, we would like to point out the importance of developing methods for detecting meaning shifts (and NLP in general), that can be applied to multiple languages. In this work, we find that a purely token-based approach already has limitations for Dutch. By not performing any lemmatization or compound splitting, we miss out on morphological variation and compounds. For morphologically rich languages, this problem might be exacerbated. From a more practical perspective, one should also consider languages with fewer digital language resources than English. The use of a large, regularly updated dictionary as we do here, or the use of huge corpora, as in Hamilton et al. (2016), might not be portable to many other languages.

7. Conclusion

We ran experiments with diachronically trained word embeddings on Dutch newswire data to see whether embeddings' similarity across years for the same term might be a reliable indicator of a meaning shift in time, as it had been at least partially observed for Italian. Rather than finding confirmation of this hypothesis, as only a couple of cases do exhibit the desired behaviour, our experiments uncovered a wide range of interfering factors that we felt required in-depth investigation. Indeed, they constitute a set of aspects that researchers in this field must take into account when designing experiments on the automatic detection of meaning shifts as well as when analysing results. With the detection and discussion of such factors in this paper, like frequency, polysemy, but also representation choices, directly linked to the results that we observe on our data, we believe that we are making an important contribution towards more robust and hopefully shared practices in the design of experiments to detect meaning shifts based on distributed representations. Finally, we also

make a practical contribution to the development of this field by making our code and our results publicly available: `https://github.com/hslh/diachronic-meaning-clin`.

Acknowledgements

We want to thank the editors of the Van Dale dictionary for providing us with valuable material regarding the terms we could use for this study. We are also grateful to the audience at CLIN 27 in Leuven for ideas, insights, and a fruitful discussion. Reviewers of this paper provided useful comments as well. Lastly, we thank Marco Del Tredici for providing the original scripts used for the Italian experiments.

References

Baroni, Marco and Roberto Zamparelli (2010), Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1183–1193. http://dl.acm.org/citation.cfm?id=1870658.1870773.

Booij, G. E. and Ariane van Santen (1998), *Morfologie: de woordstructuur van het Nederlands*, Amsterdam University Press.

Cheng, Jianpeng and Dimitri Kartsaklis (2015), Syntax-aware multi-sense word embeddings for deep compositional models of meaning, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, pp. 1531–1542. http://www.aclweb.org/anthology/D15-1177.

Del Tredici, Marco, Malvina Nissim, and Andrea Zaninello (2016), Tracing metaphors in time through self-distance in vector spaces, *Proceedings of the Third Italian Conference on Computational Linguistics*.

Evang, Kilian, Valerio Basile, Grzegorz Chrupała, and Johan Bos (2013), Elephant: Sequence labeling for word and sentence segmentation, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Seattle, Washington, USA, pp. 1422–1426. http://www.aclweb.org/anthology/D13-1146.

Frermann, Lea and Mirella Lapata (2016), A Bayesian model of diachronic meaning change, *Transactions of the Association for Computational Linguistics* **4**, pp. 31–45.

Gulordava, Kristina and Marco Baroni (2011), A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus, *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Association for Computational Linguistics, pp. 67–71.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky (2016), Diachronic word embeddings reveal statistical laws of semantic change, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pp. 1489–1501. http://aclweb.org/anthology/P16-1141.

Haser, Verena (2003), Metaphor in semantic change, *in* Barcelona, Antonio, editor, *Metaphor and Metonymy at the Crossroads, Berlin/New York: Mouton de Gruyter*, pp. 171–194.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli (2015), Sensembed: Learning sense embeddings for word and relational similarity., *ACL (1)*, pp. 95–105.

Jatowt, Adam and Kevin Duh (2014), A framework for analyzing semantic change of words across time, *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, IEEE Press, pp. 229–238.

Kenter, Tom, Melvin Wevers, Pim Huijnen, and Maarten de Rijke (2015), Ad hoc monitoring of vocabulary shifts over time, *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, ACM, New York, NY, USA, pp. 1191–1200. http://doi.acm.org/10.1145/2806416.2806474.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov (2014), Temporal analysis of language through neural language models, *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, Association for Computational Linguistics, pp. 61–65.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013), Efficient estimation of word representations in vector space, *CoRR*. http://arxiv.org/abs/1301.3781.

Řehůřek, Radim and Petr Sojka (2010), Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50.

van Noord, Gertjan, Gosse Bouma, Frank van Eynde, Daniel de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste (2013), Large scale syntactic annotation of written Dutch: Lassy, *in* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Springer, Berlin, Heidelberg, chapter 9, pp. 147–164.

Vossen, Piek, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke (2013), Cornetto: a combinatorial lexical semantic database for Dutch, *in* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Springer, Berlin, chapter 10, pp. 165–184.

Zipf, G. K. (1949), *Human Behaviour and the Principles of Least Effort*, Addison Wesley, Cambridge, MA.